Why are Children Overconfident? Developmental Differences in the Implementation of Accessibility Cues when Judging Concept Learning

Van Loon, M. H., de Bruin, A., Leppink, J., & Roebers, C. M.

(2017)

Journal of Experimental Child Psychology, 158, 77-94.

https://doi.org/10.1016/j.jecp.2017.01.008

source: https://doi.org/10.7892/boris.101010 | downloaded: 5.5.2023

Abstract

Children are often overconfident when monitoring their learning, which is harmful for effective control and learning. The present study investigated children's (N = 167, age range, 7–12 years) judgments of learning (JOLs) when studying difficult concepts. The main aims were (a) to investigate how JOL accuracy is affected by accessibility cues and (b) to investigate developmental changes in implementing accessibility cues in JOLs. After studying different concepts, children were asked to generate novel sentences, then to make JOLs, select concepts for restudy, and take a final test. Overconfidence for incorrect and incomplete test responses was reduced for older in comparison to younger children. For older age groups, generating a sentence led to greater overconfidence compared to not being able to generate a sentence, which indicates that older children relied more on accessibility cues when making JOLs. This pattern differed in the youngest age group; younger children were generally overconfident regardless of whether they had generated sentences or not. Overconfidence was disadvantageous for effective control of learning for all age groups. These findings imply that instructions to encourage children to avoid metacognitive illusions need to be adapted to children's developmental stage.

Keywords: metacognition, development, overconfidence, monitoring, control, cue utilization

Why are Children Overconfident? Developmental Differences in the Implementation of Accessibility Cues when Judging Concept Learning

Children in elementary school must develop skills that allow them to adapt to increasing study demands, including making plans, prioritizing study tasks, allocating study time, and making use of appropriate study strategies (Blair & Raver, 2015). To effectively meet these demands, it is important for children to be able to accurately monitor learning, detect errors, and identify the material that has yet to be learned (Krebs & Roebers, 2010; Roebers, Krebs, & Roderer, 2014). However, children's judgments of learning (JOLs) are often inaccurate (Finn & Metcalfe, 2014, Lipko, Dunlosky, Lipowski, & Merriman, 2012) in that most children are overconfident. That is, they are overly optimistic about their abilities, overestimate their actual performance, and often have a hard time acknowledging their errors (De Bruin & Van Gog, 2012; Finn & Metcalfe, 2014; Lipko et al., 2012). Although some overconfidence may improve motivation and task persistence (Shin, Bjorklund, & Beck, 2007), extensive overconfidence has harmful effects on learning (Dunlosky & Rawson, 2012). Typically, overconfident learners prematurely stop studying materials that they believe they know already. Hence, overconfidence can lead to ineffective self-regulation and ultimately to underachievement (Destan & Roebers, 2015; Dunlosky & Rawson 2012).

In the present study, we investigate overconfidence in elementary school children. Specifically, we aim to explain developmental differences in overconfidence, by investigating the cues that children of different ages (3rd–6th grade) use to make their judgments of learning (JOLs). To motivate the hypotheses and predictions for the age groups under investigation, we first describe findings from the literature regarding adults' overconfidence.

Explaining Overconfidence

Research with adults shows that they do not have direct access to their memory when making JOLs. Instead, they make JOLs based on a variety of cues (Brunswik, 1956; Benjamin & Bjork, 1996; Koriat, 1993, 1997), such as the perceived ease of information processing (Koriat, Ackerman, Lockl, & Schneider, 2009), the perceived familiarity with the topic of study (Griffin, Jee, & Wiley, 2009), or even the font size of studied materials (Mueller, Dunlosky, Tauber, & Rhodes, 2014). When individuals base their judgments on valid cues, monitoring is typically relatively accurate, which leads to efficiently controlled actions. However, when JOLs are based on cues that are not indicative of actual learning, a discrepancy will occur between JOLs and actual performance, leading to inaccurate judgments (Koriat, 1997).

One cue that adults based JOLs on is the accessibility of information (Koriat, 1993, 1995; Koriat & Levy-Sadot, 2001), with their JOLs tending to increase when accessibility increases (i.e. the easier and faster an individual can retrieve information, the more confident he or she will be). When accessibility is predictive of successful recall, it is considered a valid or diagnostic cue. In contrast, when accessibility is not predictive of later recall performance, the cue is considered invalid or nondiagnostic, and hence using these invalid accessibility cues will lead to inaccurate JOLs. Unfortunately, accessibility might not always be a valid cue to predict performance; in fact, accessible information may be blatantly false and not be predictive of final test performance (Benjamin, Bjork, & Schwartz, 1998; Finn & Metcalfe, 2014; Koriat, 1997).

In children, making use of invalid cues may contribute to their overconfidence. However, to date, few researchers have investigated cues and their impact on children's judgments. When judging learning, children typically implement memorability cues (Ghetti, Papini, & Angelini, 2006) and easily learned, easily remembered cues (Koriat et al., 2009). Further, children seem to

use accessibility as a cue when making JOLs; they distinguish items for which they have accessible information from those for which they do not have accessible information in memory (Koriat & Shitzer-Reichert, 2002; Schneider, Visé, Lockl, & Nelson, 2000; Van Loon et al., 2013a). As with adults, children have difficulty monitoring the quality of memory, and children's JOLs also tend to be particularly overconfident when they hold incorrect knowledge (Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013a). Further, children are typically overconfident when they make JOLs for *incomplete* test responses, meaning that they expect to receive full credit for partially correct performance (Lipko, Dunlosky, Hartwig, Rawson, Swan, & Cook, 2009; Van Loon et al., 2013a). Van Loon et al. (2013a) showed that 3rd and 4th graders accurately predicted (by correspondingly low JOLs) that performance would be low when they were not able to provide any test response (i.e., an omission error). However, when inaccurate information was accessed prior to making their JOLs, children were overconfident. For most of their incorrect test responses (i.e., commission errors), children expected full credit (i.e., they judged their commission errors were correct), despite their answers being incorrect. Furthermore, children could not distinguish partially correct responses from fully correct responses, indicating overconfidence when the information in memory was incomplete.

Accessibility seems to be particularly disadvantageous for monitoring accuracy when learners self-generate false or partially false information that serves as a cue (Castel et al., 2013; Rhodes & Castel, 2008). In such cases, generation leads to highly accessible (but nondiagnostic) cues that are then incorporated in the JOLs (Benjamin et al., 1998). In fact, it seems that JOLs are sensitive to any form of generation, regardless of whether the accessibility cue is valid or invalid (Castel et al., 2013). Adult learners report that they regularly use example generation techniques, such as keyword and sentence generation, when studying (Rawson & Dunlosky,

2016). Based on findings that generating examples can lead to improvements in learners' ability to apply learned concepts and enhance memory for concept meanings, generation has been proposed as a promising technique to support learning across different age groups (Ozubko, Hourihan, & MacLeod, 2012; Ozuru, Dempsey, & McNamara, 2009; Rawson & Dunlosky, 2016). However, although generating high-quality examples may benefit memory, the same may not be true when generating low-quality examples. Thus, generating low-quality information may be harmful for JOL accuracy, because it leads to overconfidence (Rawson & Dunlosky, 2016; Zamary, Rawson, & Dunlosky, 2016).

Children are also able to generate examples, such as when they are asked to generate example sentences (Van Loon, De Bruin, Van Gog, & Van Merriënboer, 2013b) or summaries (De Bruin, Thiede, Camp, & Redford, 2011). When generating sentences relevant to studied information, children seem to adapt their JOLs to the effort needed when generating a sentence (Van Loon et al., 2013b); that is, as they more easily generate sentences, their JOLs increase. However, it remains unclear whether self-generated information is related to overconfidence. In elementary school children, sentence generation tasks (Van Loon et al., 2013b) as well as keyword generation tasks (De Bruin et al., 2011) have been shown to improve their discrimination between correct and incorrect performance. More precisely, correlations between JOLs and performance were higher after completing generation tasks. However, these correlational measures of JOL accuracy do not reflect the degree of overconfidence, nor do they show whether children can distinguish between omission and commission errors. In order to research overconfidence, measures of absolute accuracy are needed that show the magnitude of JOLs for the different test response types (omission errors, commission errors, partially correct responses, and fully correct responses). For instance, overconfidence (i.e., poor absolute

accuracy) would be indicated by high JOLs for commission errors – that is, the elevated magnitude of JOLs inaccurately represents the fact that what they have learned (and can generate) is entirely incorrect.

Developmental Differences in Overconfidence

Although overconfidence appears to be a problem across the life span (Dunlosky & Rawson, 2012), the degree of overconfidence declines from early to later childhood (Lipko, Dunlosky & Merriman, 2009; Roebers, 2002; 2014; Schneider, 2015). The most rapid decrease in overconfidence occurs in the pre-school and kindergarten years (Destan, Hembacher, Ghetti, & Roebers, 2014; Was & Al-Harty, 2015), while further development takes place between eight and twelve years of age (Van Loon et al., 2013a).

Developmental differences in overconfidence may be due to differences in cue use between younger and older elementary school children (Buratti, Allwood, & Johansson, 2014; Pillow & Pearson, 2015). When younger children make JOLs, they might not yet be able to distinguish between desires and expectations; if so, their desires – or wishful thinking – may inappropriately raise their judgments and result in overconfidence (Schneider, 1998). Older children, however, seem less prone to wishful thinking and thus may be more likely to base their JOLs on cues derived from their learning experience (De Bruin et al., 2011; Pillow & Pearson, 2015). Even young elementary school children show sensitivity to accessibility cues; they realize when they have no information accessible in memory, and demonstrate low JOLs for omission errors in the test. However, when making JOLs for entirely incorrect responses (commission errors) and incomplete test responses (partially correct responses), JOLs are less accurate in younger than in older children. Younger elementary school children, in particular, cannot yet distinguish between low-quality and high-quality information in memory. During the elementary

school years, children experience age-related improvements in their ability to monitor uncertainty when judging the quality of accessible information in memory (Schneider, Vise, Lockl, & Nelson, 2000). JOLs for commission errors and partially correct responses become better calibrated and less overconfident with age, indicating that older elementary school children are more skilled at judging uncertainty than younger ones (Schneider et al., 2000; Von der Linden & Roebers, 2006). Because older children can more consistently implement valid cues than younger children, monitoring becomes more accurate with increasing age, and as a result, self-regulation is more effective (Destan et al., 2014; Roderer & Roebers, 2010; Schneider & Lockl, 2008).

Present Study

Although studies with a variety of learning materials demonstrated overconfidence in children and indicated that overconfidence decreases as children get older, limited research has addressed why children's JOLs improve with age. For the present study, we presume that age differences in overconfidence are due to older children being more likely to use accessibility cues that arise when children are generating sentences for concepts. To address these issues, elementary school children (grade 3 – 6) studied conceptual materials, and after study, they attempted to generate novel sentences that comprised examples of the studied concepts. Then, they made a JOL for each studied concept, with which they predicted the likelihood of correctly recalling the meaning of the studied concepts on the test. After making JOLs, they selected which concepts they would like to restudy if they had a chance, and finally, the children took a test to assess their concept learning. With this research design, we addressed whether the accessibility cues from children's generation of example sentences are related to their overconfidence. That is, we analyzed JOL accuracy for fully correct, partially correct, and

incorrect responses (i.e., omissions and commission errors) on the final test, and we investigated the relation between children's sentence generation and their JOL accuracy. Generated sentences were expected to be high in quality when participants later correctly recalled test responses. In contrast, when information in memory is incorrect or incomplete, it is likely that learners will generate low-quality sentences; basing JOLs on low-quality sentences, in turn, may lead to overconfidence.

Research on cue use mainly relies on paired-associates learning tasks. Because these studies control for prior knowledge, their results may not generalize to conceptual materials (e.g., used in classroom settings) in which prior knowledge may affect JOLs and learning (Moos & Azevedo, 2008; Toth, Daniels & Solinger, 2011). It is therefore important to account for learners' prior knowledge concerning the studied materials. Thus, in the present study, children's prior knowledge was also assessed and controlled for in the analyses. Because the aim was to address children's awareness of their errors and their incomplete or partially correct responses, difficult materials were used to ensure a sufficient number of omissions, commission errors, and partially correct responses for each learner.

First, based on previous research, we hypothesized that children would be more overconfident for commission errors than for omissions, and that they would not distinguish partially correct from fully correct responses (Hypothesis 1a). Further, we expected overconfidence to decline in the older age groups; older children were presumed to show more accurate JOLs for commission errors and partially correct responses than younger children (Hypothesis 1b).

Second, we addressed the effects of overconfidence in children's restudy selections. We expected to confirm previous research findings that children prematurely discard the items for

which they are overconfident (Hypothesis 2a), and that younger children make less adaptive restudy selections than older children (Hypothesis 2b).

The novelty of this study was testing the hypothesis that sentence generation affects children's JOLs, and that this effect is different for children in younger age groups than for those in the older age groups. Sentences generated for concepts in which later test responses result in commission errors or partially correct responses are likely to have low quality. When children make use of this self-generated information as a cue for JOLs, they are expected to be more overconfident for commission errors and partially correct responses in comparison to situations when they were not able to self-generate information (Hypothesis 3a). Finally, we hypothesized that older children were more likely to base their JOLs on accessibility of self-generated information than younger children; that is, they were expected to be more overconfident after generating a sentence in comparison to when they were not able to generate a sentence. Younger children were expected to be generally more overconfident (Hypothesis 3b).

Method

Participants and Design

The sample included 167 children from an elementary school in The Netherlands. Among these, 45 were third graders (*M* age = 8.07 years, *SD* = .62), 45 were fourth graders (*M* age = 9.16 years, *SD* = .42), 36 were fifth graders (*M* age = 10.03 years, *SD* = .56), and 41 were sixth graders (*M* age = 11.17 years, *SD* = .50). Most of the children came from middle-class families, and all of them had sufficient Dutch language abilities to understand the instructions.

The study had a mixed design, with age group as the between-subject variable and sentence generation (no sentence vs. sentence) and test response type (omission, commission error, partially correct response, fully correct response) as within-subject variables. JOLs were

analyzed for each participant as a function of their sentence generation and test response type. Participants could thus experience both levels of the sentence generation variable (being able vs. not being able to generate a sentence), in order to enable them to contrast these when making JOLs. By allowing participants to experience both levels of the sentence generation variable, their beliefs about the relation between sentence generation and test performance can be manifested (Dunlosky, Mueller & Thiede, 2016).

Materials and Procedure

As part of the task, children had to study 20 difficult concepts; these were the same concepts that were used in the study by Van Loon et al. (2013a). One concept task was developed for the 3rd and 4th graders, while another one addressed 5th and 6th graders. This was to adapt difficulty to the age group by means of the lexical richness measure (Vermeer, 2000). The concepts were largely unknown to the participants. Examples of studied concepts are shown in Appendix 1.

The task consisted of six subtasks: 1) pre-test, 2) concept study, 3) sentence generation, 4) JOLs, 5) restudy selection, and 6) concept learning test. All subtasks were presented in a booklet, with blank pages between the subtasks. The order of the concepts was randomized across children and across subtasks; children received different versions of the concept tasks, and the order of concepts was different for every subtask.

Children were tested in the classroom, the session lasted approximately one hour. The experimenter and the teacher supervised the participants to ensure that they were working on the right task and not rechecking their booklet. The experimenter informed children that they would study concepts and that their understanding of these concepts would be checked on the final test. Participants were informed that test responses that were paraphrases of the original studied

meaning would be marked as correct. Further, the children were told to wait for the experimenter to announce when blank pages could be turned over. Before starting, participants were provided with an example of a concept and a test question. Then, participants received the booklet and began completing the pre-test.

The *pretest* assessed children's prior knowledge. It consisted of a list of the concepts to be studied on one page, containing a space to write down responses about the meaning next to each concept. Participants were instructed to fill in the meaning if they thought they knew the concept, and to leave the space blank if they did not.

After 10 minutes, children completed the *concept study task*, which consisted of 20 concepts. Concepts were presented with their meaning and an example sentence. Appendix 1 shows examples of the concepts that had been studied. A pilot study revealed that 10 minutes of study time was needed in order for children to read concepts thoroughly, yet not become inattentive.

After studying, children completed the *sentence generation* task, and access to the definitions was no longer permitted. The concepts were listed on one page, including spaces to write down a sentence for each one. The experimenter referred to the example sentences children had seen during study, and instructed them to make their own example sentences by correctly using the studied concepts. Participants were encouraged to generate novel, sensible sentences for every concept, and were instructed to leave the space blank, if they thought they would not be able to do so.

For the *JOL* task, the studied concepts were presented next to a thermometer. JOLs were provided by marking a square on an 11-point thermometer (this method has been used for the same age groups by Van Loon et al., 2013a), ranging from "certain I will not recall the meaning"

to "certain I will recall the meaning". Before they were required to make JOLs, children were shown the example concept again, and the JOL thermometer was explained. The experimenter explained what the various points on the thermometer meant in terms of likelihood to recall the meaning of a concept on the test. Then, participants made JOLs for each concept by marking the corresponding square of choice on the 11-point thermometer.

For the *restudy selections*, the concepts were displayed on a page in a 10 × 2 array, and participants were asked to mark concepts they would like to restudy. In addition, children were reminded that they would be tested on the studied concepts. They were instructed to mark those concepts they wished to restudy before taking the test. Participants were not given the opportunity to restudy the selected concepts, because an additional study trial would interfere with the relation between JOLs and test performance.

During the *concept test*, the studied concepts were presented to the participants, and they were asked to write down their meanings. The test took approximately 15 minutes.

Children did not receive feedback on their responses in any phase of the study. Upon completion, participants were thanked for their participation, and they received a small gift.

Scoring

Appendix 1 shows examples of studied concepts, provided sentences, and test responses as well as the corresponding scoring. Pre- and post-test responses were scored as either omission (no response given), commission error (an entirely incorrect response given, a response which does not contain any idea unit of the concept meaning), partially correct (response contains some, but not all, units of the concept meaning), or fully correct (response contains all correct idea units). Similar to research on adults' concept learning (Rawson & Dunlosky, 2007), the meaning of each studied concept consisted of 2 to 4 idea units (M = 3.0, SD = .64). In line with

the scoring used by Rawson and Dunlosky (2007), idea units were scored as correct when these were either provided verbatim or as a paraphrase of the idea unit. Sentences were scored as "no sentence", "incorrect sentence" (the sentence does not use the word in a meaningful way), and "correct sentence" (the studied word is used correctly in the generated sentence). To assess interrater reliability, two independent raters scored 27% of pre-test performance (Kappa = .86; 23.4% of performance on the sentence generation task, Kappa = .86; 22% of the post-test responses, Kappa = .79; showing good reliability of scoring).

Analyses

We investigated JOL magnitudes and restudy selections for commission errors and partially correct responses on the final test. Further, we also examined the influence of Grade Level and Sentence Generation. In addition, the factor Prior Knowledge was added to the analyses to account and control for potential effects. Moreover, we explored how the quality of generated sentences is related to performance and JOLs. For all analyses, SPSS version 23 was used, including Bonferroni corrections for post-hoc tests. Differences between test response types and age groups are only reported when significant at $p < .05$.

Firstly, general linear model (GLM) analyses for repeated measures were used to investigate differences in mean prior knowledge, test performance, JOL magnitudes, and restudy; Test Response Type (omission, commission error, partially correct response, fully correct response) was entered as a within-subject repeated factor, and Grade Level (Grade 3, 4, 5, or 6) was entered as a between-subject factor. Significant interaction effects between Test Response Type and Grade Level are followed up with a multivariate GLM. For significant effects, partial eta square (η_p^2) is reported as the measure of effect size.

Because our hypotheses specifically address underlying reasons for overconfidence in commission errors and partially correct responses, the focus is placed on these two response types when testing our hypotheses on overconfidence. A linear mixed model analysis was used to analyze the effects of Grade Level (3-6), Test Response Type (commission errors, partially correct responses), Sentence Generation (sentence or no sentence), and Prior Knowledge (prior knowledge response or no prior knowledge response) on JOLs. This type of multi-level modeling is appropriate for these data because of its nested structure (the response types are nested under each participant). With the mixed model, we can address our hypotheses concerning the effects of item-level predictors (generation of a sentence), take participants' prior knowledge into account, and investigate the effects of a participant-level predictor (Age Group) on an item-level dependent variable (JOLs). Collapsing across items would reduce power as well as variability in the item-level predictors, whereas treating items independently of participants would lead to overpowered and inappropriate modeling of the variability shared among items from the same participant. Individual differences in learning ability, prior knowledge, generated sentences, JOL distributions, restudy selections, and test outcomes are expected. Therefore, a random intercept is used for each participant to account for variance in individual observations. Items are entered as repeated measurement, and for significant effects, the estimated value of the unstandardized regression coefficient (*b*) is reported as a measure of effect size, to show the change in JOL magnitude.

Results

First, we evaluated to what extent children are overconfident concerning their commission errors and partially correct responses, whether overconfidence declines in older age groups, and how overconfidence affects restudy selections. Second, we investigated how self-generation of

sentences is related to overconfidence and whether sentence generation affects JOLs differently in younger than in older children.

Before testing our hypotheses, we present descriptive statistics on prior knowledge and test performance. Table 1 shows prior knowledge in the pre-test, and performance in the post-test for all of the four age groups. As shown in Table 1, children had low prior knowledge. The GLM analysis on percentage of prior knowledge (overall percentage of generated prior knowledge and percentage of correct prior knowledge as repeated measurements) showed a significant effect for Age Group ($F(3, 163) = 10.95$, $p < .001$, $\eta_p^2 = .17$; 3rd grade < 4th and 6th grade (ps < .001); 4th grade > 5th grade; 5th grade < 6th grade (ps < .05)). As two age-adapted task versions were used, one with materials for grades 3 and 4, and one for grades 5 and 6, the observation that grades 3 and 5, and grades 4 and 6, respectively, did not differ in prior knowledge indicates that the two versions were comparable in difficulty.

For post-test performance, there was a significant interaction between Age Group and Test Response Type ($F(9, 489) = 14.96$, $p < .001$, $\eta_p^2 = .22$). Multivariate analyses showed that the percentage of omissions differed between Age Groups ($F(3, 163) = 15.46$, $p < .001$, $\eta_p^2 = .22$; (3rd grade > 4th, 5th , and 6th grades, all ps < .01)), as did the percentage of commission errors ($F(3, 163) = 6.14$, $p = .001$, $\eta_p^2 = .10$; 3rd grade > 6th grade ($p = .002$); 4th grade > 6th grade ($p = .005$)). Further, Age Group affected the percentage of partially correct responses ($F(3, 163) = 9.8$, $p < .001$, $\eta_p^2 = .15$; 3rd grade < 4th, 5th, and 6th grades (all ps < .01)). Finally, the number of correct responses increased for the older age groups ($F(3, 163) = 11.12$, $p < .001$, $\eta_p^2 = .29$ (3rd grade < 4th, 5th, and 6th grades; 4th grade < 5th and 6th grades, all ps < .05)).

Overconfidence and Effects of Age Group

JOLs were made on an 11-point scale, ranging from 0 to 10 points. Table 1 shows the mean JOL magnitudes for the four age groups and the four test response types. Figure 1 shows the mean JOLs as a function of test responses.

--- Insert Table 1 approximately here ---

As Figure 1 shows, JOLs were disproportionally high for commission errors; accurate JOLs for commission errors should have been as low as JOLs for omissions (because children did not receive credit for these responses). Further, for the partially correct responses, children's JOLs were nearly as high as JOLs for fully correct responses.

--- Insert Figure 1 approximately here ---

The effect of Response Type ($F(3, 85) = 374.25$, $p < .001$, $\eta_p^2 = .93$) indicates that JOLs for the four test response types differed from each other. JOLs for omissions ($M = 2.06$) were lower than JOLs for commission errors ($M = 7.14$), partially correct responses ($M = 8.74$), and fully correct responses ($M = 9.00$), all ps $< .001$. JOLs for commission errors were lower than JOLs for partially correct responses and fully correct responses (ps $< .001$). However, there was no significant difference between JOLs for partially correct and fully correct responses. The finding that children give higher JOLs for commission errors than for omissions, and do not differentiate between partially and fully correct responses, confirms their overconfidence (Hypothesis 1a).

Age Group significantly affected mean JOLs ($F(3, 87) = 3.1$, $p = .031$, $\eta_p^2 = .10$). Overall, mean JOLs did not differ between grade 3 ($M = 7.38$) and grade 4 (M magnitude = 6.80), but differed marginally between grade 3 and grade 5 (M magnitude = 6.40, $p = .055$) as well as between grade 3 and grade 6 (M magnitude = 6.36, $p = .055$). There were no other significant differences between age groups. JOLs for grade 3 were highest, but in contrast, their performance was lowest, which confirms that the youngest children were the most overconfident when judging learning (Hypothesis 1b).

--- Insert Figure 2 approximately here ---

Figure 2 shows the effects of Age Group on JOLs for commission errors (Figure 2a) and partially correct responses (Figure 2b), and shows that JOLs were lower for older than for younger children. The linear mixed model analysis shows that Age Group significantly affected overconfidence for commission errors and partially correct responses ($b = .80$, $SE = .32$, $F(3, 333.98) = 25.20$, $p < .001$). Overconfidence declined with age; JOLs for 3rd grade > 4th, 5th, and 6th grades (all ps < .001). JOLs for 4th grade > 6th grade ($p = .004$), which further confirms that JOLs were lower and more accurate for the older age groups than for the younger age groups (Hypothesis 1b).

In addition, prior knowledge generation significantly affected JOLs ($b = .56$, $SE = .26$, $F(1, 845.06) = 17.72$, $p < .001$). Regardless of whether prior knowledge was correct, it served as a cue for the JOLs; children were more overconfident when they generated a prior knowledge response before study than when they were not able to generate any prior knowledge. There was no

significant interaction between prior knowledge generation and Age Group (p = .860), indicating similar effects of prior knowledge on JOLs for all children.

Relation between JOLs and Restudy Selections

To appropriately select concepts for restudy, the children should select their omissions, commission errors, and partially correct responses. Response Type was related to Restudy Selections ($F(3, 258) = 209.83$, $p < .001$, $\eta_p^2 = .71$); the percentage of restudy selections for omissions was higher than the percentage for commission errors, partially correct responses, and fully correct responses (all ps < .001). Further, selections for commission errors were higher than fully correct responses ($p < .001$), and partially correct responses were more often selected than fully correct responses ($p = .05$). Table 1 shows the percentage of restudy selections for the four test response types and the four grades. While children selected most of their omissions for further study, Table 1 shows that this was not the case for commission errors and partially correct responses, confirming that overconfidence is related to maladaptive restudy selections (Hypothesis 2a). However, there was no effect of Age Group on restudy selections ($F(3, 86) = .37$, $p = .778$) (contrasting Hypothesis 2b).

To investigate the relation between monitoring and control, intra-individual Goodman–Kruskal gamma correlations (Thiede & Dunlosky, 1999) were calculated between JOLs and restudy selections (1 = selected, 0 = not selected). Mean gammas were -.93 (SD = .16) for grade 3, -.76 (SD = .54) for grade 4, -.84 (SD = .40) for grade 5, and -.85 (SD = .39) for grade 6. These negative correlations show that selections are strongly related to JOLs: children were more likely to restudy concepts for which they gave low JOLs. A univariate analysis shows that gamma correlations between JOLs and restudy selections were not affected by Age Group ($F(1, 129) < 1$). Overconfidence thus caused children to prematurely discard concepts for further study. In

addition, restudy selections were maladaptive for commission errors and partially correct responses for all age groups.

Effects of Sentence Generation Cues on Overconfidence

Overall, sentences were generated for 47.4% (*SD* = 21) of the studied concepts, with mean percentages of 32.8% (*SD* = 14) for 3rd grade, 51.1% (*SD* = 25) for 4th grade, 49.2% (*SD* = 18) for 5th grade, and 57.9% (*SD* = 18) for 6th grade, respectively. Age Group affected the mean number of generated sentences ($F(3, 163) = 13.21$, $p < .001$, $\eta_p^2 = .20$; 3rd grade < 4th grade, 5th grade, and 6th grade (all *p*s < .001)).

In order to investigate the types of accessibility cues children used, the quality of generated sentences and the relationship between the quality of sentences and test performance was assessed. While 76.2% of the generated sentences were rated as correct, the remaining sentences were scored as incorrect. For 75.1% of the correct sentences, the word was placed in a novel example sentence. This was not the case for 2.8% of the remaining correct sentences, but these sentences still showed that the child had understood the concept. For the remaining 22.1% of correct sentences, children described the definition of the concept. For the incorrect sentences, children did not show any correct and sensible usage of the studied concept. Appendix 1 shows examples of sentences and their scoring.

Test performance was higher for correct sentences (*M* performance = .59; *SD* = .37) compared to incorrect ones (*M* = .27; *SD* = .34). Gamma correlations between sentence quality and performance were calculated to investigate whether sentence quality was a predictor of performance, and gamma correlations between sentence quality and JOLs were calculated to investigate children's cue utilization of sentence quality. Gammas were calculated for each age group, because individual correlations per participant would have led to a high number of

missing data points (in order to calculate gamma, a sufficient number of correct and incorrect sentences are needed, and further, there should be variance in JOLs and in test performance for these correct and incorrect sentences to prevent ties). The moderate to strong correlations between sentence quality and performance, that is, Gamma = .63 for 3rd grade, .61 for 4th grade, .74 for 5th grade, and .59 for 6th grade, gave evidence that sentence quality was predictive of subsequent test performance. When children generated correct sentences, 42.8% of their following test responses were partially correct, and 37.9% of their responses were fully correct. Only 6.8% of the responses following correct sentences were omissions, and 12.4% of responses were commission errors. In contrast, test responses following incorrect sentences resulted in 45.8% commission errors and 27.6% omissions; only 6.3% were fully correct, whereas 20.3% were partially correct.

--- Insert Figure 3 approximately here ---

Figure 3 shows the effect for generation of correct vs. incorrect sentences on JOLs; JOLs were typically high after sentence generation; regardless of their correctness. Accurate JOLs should be based on valid cues, meaning that children should be aware of the sentence quality and consistently give lower JOLs for sentences with low quality. The gamma correlations between sentences (correct or incorrect) and JOLs were .42 for 3rd grade, .31 for 4th grade, .51 for 5th grade, and .46 for 6th grade, and show that children in all age groups moderately differentiated between incorrect and correct sentences. However, as Figure 3 shows, JOLs for incorrect sentences were high, even when their sentences were not indicative of final test performance.

Sentence generation significantly affected overconfidence for commission errors and partially correct responses (b = 3.31, SE = .41, F(1, 1071.05) = 204.13, p < .001). Mean JOLs were as high as 8.87 (SD = 2.2) points when a sentence was generated and only 6.10 points (SD = 3.9) when no sentence was generated, confirming Hypothesis 3a.

--- Insert Figure 4 approximately here ---

Importantly, there was a significant interaction effect between Sentence Generation and Age Group (F(3, 877.93) = 9.67, p < .001); this interaction effect is depicted in Figure 4. Although the younger children discriminated in their JOLs between sentence generation and no sentence generation, the older children did so more strongly. This was made clear by a larger difference between JOLs for concepts for which children did not generate sentences compared to concepts for which they did. Follow-up tests investigating the effects of sentence generation on overconfidence for all four grades, separately, show that all age groups significantly discriminated between generating a sentence and not generating a sentence: third grade, b = .87, SE = .22, F(1, 125.44) = 15.84, p < .001; fourth grade, b = 3.38, SE = .29, F(1, 178.21) = 138.72, p < .001; fifth grade, b = 4.61, SE = .46, F(1, 104.09) = 101.34, p < .001; and sixth grade, b = 3.93, SE = .42, F(1, 208.62) = 86.82, p < .001. However, the effect size indicated by the b value of the difference between sentence generation and no sentence generation was larger for the older children than for the younger children. This is particularly pronounced when comparing third and fourth grade. Thus, younger children, and especially third graders, were more overconfident when they did not generate a sentence than older children. Further, they did not discriminate strongly between having generated a sentence or not. The finding that older

children's JOLs were more sensitive to accessibility cues, whereas younger children were generally more overconfident, confirming Hypothesis 3b.

Discussion

The present study sought to evaluate and explain age-related differences in overconfidence for incorrect (commission errors) and incomplete (partially correct) test responses in children (third to sixth grade) when studying difficult concepts. As expected, children from third to sixth grade were overconfident when judging learning of difficult concepts; these findings confirm previous research on young learners' overconfidence (Lipko-Speed, 2013; Roebers, 2014; Schneider & Lockl, 2008). Children were somewhat overconfident for test omissions, but, what is more concerning, they were highly overconfident when making JOLs for commission errors and partially correct test responses. In fact, when monitoring learning, the children did not distinguish between partially correct and fully correct responses.

In particular, the youngest children (third graders) were disproportionally biased for incorrect and incomplete test responses compared to the older age groups. For 4th, 5th, and 6th graders, differences in overconfidence were smaller and less pronounced. This finding confirmed recent observations by Was and Al-Harthy (2015), who reported that in classroom settings, overconfidence only begins to decrease in 4th graders. Moreover, our results indicate a slowdown in the developmental progression in monitoring accuracy after 4th grade. However, despite the level of overconfidence being lower for older children, their JOLs for commission errors and partially correct responses were not accurate, either. This was especially the case after generating a sentence prior to making JOLs. We will return to this finding below.

Most previous research on children's monitoring accuracy used paired-associates learning tasks and tested children individually. The present study, however, extended results on

overconfidence to the classroom, for which groups of children were tested with a more complex and more educationally valid learning task. There has been debate about whether overconfidence is harmful for children; the adaptivity hypothesis states that a slight degree of overconfidence can be beneficial for learning because it motivates children to persist in a task (Shin et al., 2007; Bjorklund, Periss, & Causey, 2009). However, in educational settings, children's overconfidence is likely to be harmful as it leads to ineffective control of study (Dunlosky & Rawson, 2012). Hence, overestimators have been shown to have lower performance than children who judge their learning more accurately (Destan & Roebers, 2015). This study also demonstrates that overconfidence is harmful for effective control by showing that it was detrimental for the effectiveness of restudy selections across different age groups. Findings on the relation between JOLs and restudy selections show that children did not select the items for which they were overconfident. That is, they decided to discard most of the concepts from further study if their test responses were commission errors or only partially correct. Contrary to our expectations, restudy selections did not improve in older children. Non-adaptive restudy selections do not seem to be due to the children's lack of motivation to study. Otherwise, they would not have selected most of their omission errors, for which they more accurately judged their corresponding knowledge.

Children's Cue Use when Monitoring Learning

It is important to note this study points out underlying reasons for children's overconfidence by clarifying how children's JOLs are related to accessibility cues; in this case, these cues arose as children attempted to generate (i.e., accessed) sentences for each concept. When children were able to self-generate a sentence prior to judging learning, their JOLs increased. When correct sentences were generated, it was likely that performance was also high,

so children's high JOLs were appropriate. In contrast, sentence generation was disadvantageous for monitoring accuracy when a concept had not yet been sufficiently learned. Children were generally more overconfident when they had generated sentences compared to when they were not able to generate a sentence. Although a thorough discussion of JOLs for omission errors is beyond the scope of this paper, it is still important to note that JOLs for omissions were also higher when children generated sentences than when they did not. The accessibility hypothesis (Koriat, 1993, 1995) states that people often cannot evaluate whether accessed information is correct or incorrect. Following this, judgments are mainly based on the quantity, instead of the quality of accessed information. Taken together, our findings demonstrate that children used accessibility cues for judgments, even when this cue did not give valid indications about the quality and accuracy of their knowledge.

Children in all age groups were highly overconfident when generating a sentence for commission errors and partially correct responses, although the degree of their overconfidence decreased for older age groups. Interestingly, older children were more affected by the sentence generation task when making JOLs, implying that there were age-related changes in cue utilization. When they were not able to generate a sentence, older children judged their learning more accurately than younger children did. More precisely, when not generating a sentence, sixth grade learners' JOLs were low; for the partially correct responses, their JOLs were close to the midpoint of the scale and therefore nearly accurate. In contrast, younger children were generally overconfident when judging learning for commission errors and partially correct responses; this was also the case when they were not able to generate a sentence. This may indicate that, when monitoring learning, older children more strongly rely on accessibility cues than younger children. It is probable that the older children based their JOLs on the accessibility of

information in the same way that adults do (Castel et al., 2013; Dunlosky, Rawson, & Middleton, 2005).

In contrast to older children, the younger ones do not seem to base their JOLs to the same extent on accessibility cues; they were still overconfident when not being able to generate a sentence. This finding might be explained by the wishful thinking hypothesis (Lipko, Dunlosky, & Merriman, 2009; Lipko-Speed, 2013; Schneider, 1998); younger children may base judgments of abilities on their wishes to perform on the final test, instead of on their actual expectation of how well they learned the concept. The older children seemed better able to take distinguishable cues into account (Schneider, 1998; Von Helversen, Mata, & Olsson, 2010).

Recently, Pillow and Pearson (2015) suggested that older children and adults may be more sensitive to differential features of cognitive activities (related to effort and difficulty) than younger children. The present findings support this assumption. More precisely, our results suggest that sensitivity to distinctive cues may increase with age, and show that older children are more sensitive to accessibility cues emerging from the generated sentences. Moreover, our findings imply a rapid decrease in overconfidence and a pronounced increase in implementation of accessibility cues taking place between grades 3 and 4. Apparently, when monitoring concept learning, extensive maturation in metacognition occurs around 9 and 10 years of age.

A further explanation for why the youngest age group had the most difficulties in monitoring their knowledge of concepts may be their low task performance. Literature shows a bi-directional relationship between monitoring and learning; while confidence judgments affect learning, domain-specific task knowledge also affects confidence judgments (Roebers & Spiess, 2016; Schneider, 2015). Even though task difficulty was adapted to age level, and prior knowledge was similar for the younger and older children, the youngest children (8-year-olds)

were less capable of studying the concepts. This was indicated by a lower number of generated sentences and lower test performance. Possibly, it was more difficult for younger children than for older primary school children to learn the 20 new concepts, which may be due to younger children's less advanced reading skills, a lack of cognitive resources (such as working memory capacity), and limited educational experience with concept learning. Monitoring is affected by cognitive resources (DeMarie & Ferron, 2003), hence, completing the study tasks may have been easier for the older children, which in turn has enabled them to make use of more cognitive resources when monitoring uncertainty.

In addition, prior knowledge generation was accounted for in the analyses. When children provided a prior knowledge response in the pre-test, this led to more overconfidence for commission errors and partially correct responses compared to not being able to generate prior knowledge. Thus, activation of inaccurate prior knowledge can have harmful effects on monitoring accuracy (Van Loon et al., 2013a). Our findings suggested that prior knowledge activation can lead to reliance on accessibility cues, and when this cue is invalid, it can cause overconfidence for incorrect and incomplete test responses. However, the effect size of prior knowledge activation on JOLs was smaller than the effect size of sentence generation. It is relevant to note that the children had minimal prior knowledge; they only provided prior knowledge responses for 28% of the concepts. Sentence generation response, however, was provided for 50% of the concepts. This discrepancy could be the reason why sentence generation had stronger effects on JOLs than prior knowledge activation. This assumption merits testing in future research.

Limitations and Future Directions

One limitation of the present study is that implementation of accessible cues was only measured according to self-generation of sentences. Despite the fact that self-generation of sentences has proven to be a powerful cue, when judging learning in real life, multiple cues must be taken into account. Hence, other cues that are also likely to be powerful, such as study time (Koriat et al., 2009), were not measured. Future research could address the ways in which children take multiple cues into account when judging learning.

Using a difficult task in order to address monitoring for incorrect and incomplete test performance – as has been purposely done in the present study – may also be a limitation affecting JOLs. Hence, overconfidence may be due to the hard/easy effect (Lichtenstein & Fischhoff, 1977), that is, learners are typically overconfident for tests consisting of difficult items and underconfident for easy tests. A strength of the present analysis, however, is that we did not calculate a single mean score per person indicating the level of over- or underconfidence, but instead, we compared JOL magnitudes for omissions, commission errors, partially correct responses, and fully correct responses. However, using easier materials could have helped to discriminate between correct and incorrect performance. Future research should study the effects of task difficulty on children's overconfidence.

In summary, we can conclude that overconfidence was high for commission errors and partially correct responses, and as a consequence, children decided not to further study the concepts that were not yet sufficiently learned. To our knowledge, this is the first study investigating how elementary school children's JOLs are related to accessibility cues, and how implementation of these cues is affected by child development. Although students appreciate generation strategies and endorse them (McCabe, 2011), and research also shows that generation tasks can support learning (Rawson & Dunlosky, 2016), the present study shows that these tasks

do not always help children to evaluate their actual abilities. Instead, it shows that generation can lead to overconfidence for materials that have not yet been sufficiently learned. Our findings imply that teachers should support their students to accurately monitor learning by avoiding metacognitive illusions (i.e., systematic errors in metacognitive monitoring). Addressing monitoring and control for materials that have not been sufficiently learned and incorrect test performance should be a priority in research and educational practice. Hence, by gaining insight into processes that hinder, and factors that improve, monitoring of incorrect performance, learning could be improved (Efklides, 2011; Krebs & Roebers, 2010). Future research should investigate interventions to support participants with accurate monitoring when performing a generation task. For example, asking children to compare their sentences with a correct standard may be beneficial for monitoring accuracy (Dunlosky, Hartwig, Rawson, & Lipko, 2011). Importantly, this study indicates that instructional interventions aimed at supporting children to accurately judge their learning need be adapted to children's developmental stage. Younger children seem to be generally overconfident, whereas for older children, overconfidence seems to be a result of taking accessibility cues into account that do not indicate the quality of their learning.

References

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition: The 27th Carnegie symposium on cognition* (pp. 309-338). Hillsdale, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55. doi: 10.1037/0096-3445.127.1.55

Bjorklund, D. F., Periss, V., & Causey, K. (2009). The benefits of youth. *European Journal of Developmental Psychology, 6*(1), 120-137. doi: 10.1080/17405620802602334

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*, 711-731. doi: 10.1146/annurev-psych-010814-015221

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Buratti, S., Allwood, C. M., & Johansson, M. (2013). Stability in the metamemory realism of eyewitness confidence judgments. *Cognitive Processing, 15*(1), 39-53. doi: 10.1007/s10339-013-0576-y

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: the use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41*(1), 28-35. doi: 10.3758/s13421-012-0249-6

De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school

children. *Journal of Experimental Child Psychology, 109*(3), 294-310. doi: 10.1016/j.jecp.2011.02.005

De Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245-252. doi: 10.1016/j.learninstruc.2012.01.003

DeMarie, D., & Ferron, J. (2003). Capacity, strategies, and metamemory: Tests of a three-factor model of memory development. *Journal of Experimental Child Psychology*, *84*(3), 167-193. doi: 10.1016/S0022-0965(03)00004-3

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5-to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213-228. doi: 10.1016/j.jecp.2014.04.001

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, 1-28. doi: 10.1007/s11409-014-9133-z

Dunlosky, J., Hartwig, M. K., Rawson, K., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64*(3), 467-484. doi: 10.1080/17470218.2010.502239

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271-280. doi: 10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551-565. doi: 10.1016/j.jml.2005.01.011

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S.K. Tauber (Eds), *The Oxford Handbook of Metamemory* (pp. 23-37). Oxford, UK: Oxford University Press.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6-25. doi: 10.1080/00461520.2011.538645

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction, 32*, 1-9. doi: 10.1016/j.learninstruc.2014.01.001

Ghetti, S., Papini, S., & Angelini, L. (2006). The development of the memorability-based strategy: Insight from a training study. *Journal of Experimental Child Psychology*, *94*(3), 206-228. doi: 10.1016/j.jecp.2006.01.004

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001-1013. doi: 10.3758/MC.37.7.1001

Koriat, A. (1993). How do we know that we know: The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609-639. doi: 10.1037/0033-295x.100.4.609

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*(3), 311-333. doi: 10.1037/0096-3445.124.3.311

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology-General, 126*(4), 349-370. doi: 10.1037/0096-3445.126.4.349

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The easily learned, easily remembered heuristic in children. *Cognitive Development, 24*(2), 169-182. doi: 10.1016/j.cogdev.2009.01.001

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*(1), 34-53. doi: 10.1037/0278-7393.27.1.34

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy. In P. Chambres, M. Izaute & P. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 1-17). New York: Kluwer. doi: 10.1007/978-1-4615-1099-4

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80*(3), 325-340. doi: 10.1348/000709910x485719

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?. *Organizational behavior and human performance*, *20*(2), 159-183. doi:10.1016/0030-5073(77)90001-0

Lipko-Speed, A. R. (2013). Can young children be more accurate predictors of their recall performance? *Journal of Experimental Child Psychology, 114*(2), 357-363. doi: 10.1016/j.jecp.2012.09.012

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*(4), 307-318. doi: 10.1037/a0017599

Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young Children are not Underconfident With Practice: The Benefit of Ignoring a Fallible Memory Heuristic. *Journal of Cognition and Development, 13*(2), 174-188. doi: 10.1080/15248372.2011.577760

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology, 103*, 152-166. doi: 10.1016/j.jecp.2008.10.002

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462-476. doi:10.3758/s13421-010-0035-2

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology, 33*(2), 270-298. doi: 10.1016/j.cedpsych.2007.03.001

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1-12. doi: 10.1016/j.jml.2013.09.007

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory, 20*(7), 717-727. doi: 10.1080/09658211.2012.699070

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228-242. doi: 10.1016/j.learninstruc.2008.04.003

Pillow, B. H., & Pearson, R. M. (2015). Children's and adults' judgments of the controllability of cognitive activities. *Metacognition and Learning*, 10, 231-244. doi: 10.1007/s11409-014-9122-2

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4-5), 559-579. doi: 10.1080/09541440701326022

Rawson, K. A., & Dunlosky, J. (2016). How Effective is Example Generation for Learning Declarative Concepts? *Educational Psychology Review*, *28*(3), 649-672. doi:10.1007/s10648-016-9377-z

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615. doi: 10.1037/a0013684

Roderer, T., & Roebers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition and Learning, 5*, 229-250. doi: 10.1007/s11409-010-9059-z

Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*(6), 1052. doi: 10.1037/0012-1649.38.6.1052

Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. *The Wiley handbook on the development of children's memory, Volume I/II*, 865-894. doi: 10.1002/9781118597705.ch37

Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences*, *29*, 141-149. doi: 10.1016/j.lindif.2012.12.003

Roebers, C. M., & Spiess, M. (2016). The development of metacognitive monitoring and control in second graders: A short-term longitudinal study. *Journal of Cognition and Development*. doi: 10.1080/15248372.2016.1157079

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science, 1*(2), 291-297. doi: 10.1111/1467-7687.00044

Schneider, W. (2015). *Memory development from early childhood through emerging adulthood*. New York: Springer. doi: 10.1007/978-3-319-09611-7

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (Vol. 14, pp. 391-409). Mahwah, NJ: Lawrence Erlbaum.

Schneider, W., Vise, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring - Evidence from a judgment-of-learning task. *Cognitive Development, 15*(2), 115-134. doi: 10.1016/s0885-2014(00)00024-1

Shin, H., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's overestimation in a strategic memory task. *Cognitive Development, 22*(2), 197-212. doi: 10.1016/j.cogdev.2006.10.001

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of*

Experimental Psychology: Learning, Memory and Cognition, 25(4), 1024-1037. doi: 10.1037/0278-7393.25.4.1024

Toth, J. P., Daniels, K. A., & Solinger, L. A. (2011). What you know can hurt you: Effects of age and prior knowledge on the accuracy of judgments of learning. *Psychology and Aging, 26*(4), 919-931. doi: 10.1037/a0023379

Van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (2013a). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15-25. doi: 10.1016/j.learninstruc.2012.08.005

Van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (2013b). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*(2), 173-191. doi: 10.1007/s11409-013-9100-0

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65-83. doi: 10.1177/026553220001700103

Vo, V. A., Li, R., Kornell, N., Pouget, A., & Cantlon, J. F. (2014). Young Children Bet on Their Numerical Skills: Metacognition in the Numerical Domain. *Psychological Science, 25*(9), 1712-1721. doi: 10.1177/0956797614538458

Von der Linden, N., & Roebers, C. M. (2006). Developmental changes in uncertainty monitoring during an event recall task. *Metacognition and Learning, 1*(3), 213-228. doi: 10.1007/s11409-006-9001-6

Von Helversen, B., Mata, R., & Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue abstraction processes in multiple-cue judgment. *Developmental Psychology, 46*(1), 220. doi: 10.1037/a0016690

Was, C. A., & Al-Harthy, I. (2015). Developmental differences in overconfidence: When do children understand that attempting to recall predicts memory performance? *The Researcher, 27* (1), 1 - 5.

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction*, *46*, 12-20. doi: 10.1016/j.learninstruc.2016.08.002

Table 1

Prior knowledge responses (in %); test performance (in %), JOLs (ranging from 0-10 points) and restudy selections (in %) for the test response types (omissions, commission errors, partially correct responses and fully correct responses) for the four age groups.

	Omissions	**Commission Errors**	**Partially Correct Responses**	**Fully Correct Responses**
			Prior Knowledge (%)	
Grade 3	82.5 (12)	14.16 (8)	8.55 (5)	5.00 (0)
Grade 4	66.94 (18)	18.14 (10)	15.13 (7)	6.91 (3)
Grade 5	77.50 (12)	15.48 (8)	10.22 (6)	6.33 (4)
Grade 6	69.55 (16)	16.76 (11)	14.34 (8)	6.25 (4)
			Test Performance (%)	
Grade 3	64.38 (12)	14.36 (9)	17.29 (9)	3.96 (5)
Grade 4	46.93 (20)	13.67 (13)	26.49 (13)	12.98 (12)
Grade 5	50.81 (18)	9.44 (8)	17.08 (11)	22.67 (16)
Grade 6	40.12 (18)	6.56 (8)	27.29 (13)	26.05 (19)
			JOL Magnitudes	
Grade 3	2.64 (3.4)	8.37 (3.1)	9.38 (1.6)	9.44 (1.6)
Grade 4	1.73 (2.9)	7.60 (3.2)	8.90 (2.2)	9.26 (1.6)
Grade 5	1.75 (2.5)	6.68 (3.5)	8.63 (2.3)	9.11 (1.9)
Grade 6	1.17 (2.1)	6.00 (3.6)	8.52 (2.7)	9.08 (2.0)
			Restudy Selections (%)	
Grade 3	79.54 (40)	15.94 (37)	7.36 (26)	2.27 (15)
Grade 4	78.23 (41)	21.99 (42)	15.09 (36)	11.94 (33)
Grade 5	86.07 (35)	26.47 (44)	9.30 (29)	5.95 (24)
Grade 6	85.23 (36)	28.57 (46)	12.39 (33)	8.15 (27)

Note. Standard deviations of the mean in parentheses.

Figure 1

JOLs in points (ranging from 0-10) for the test responses (omissions, commission errors, partially correct responses, and fully correct responses). The figure shows overconfidence for omissions, commission errors, and partially correct responses. Error bars indicate the 95% confidence interval.

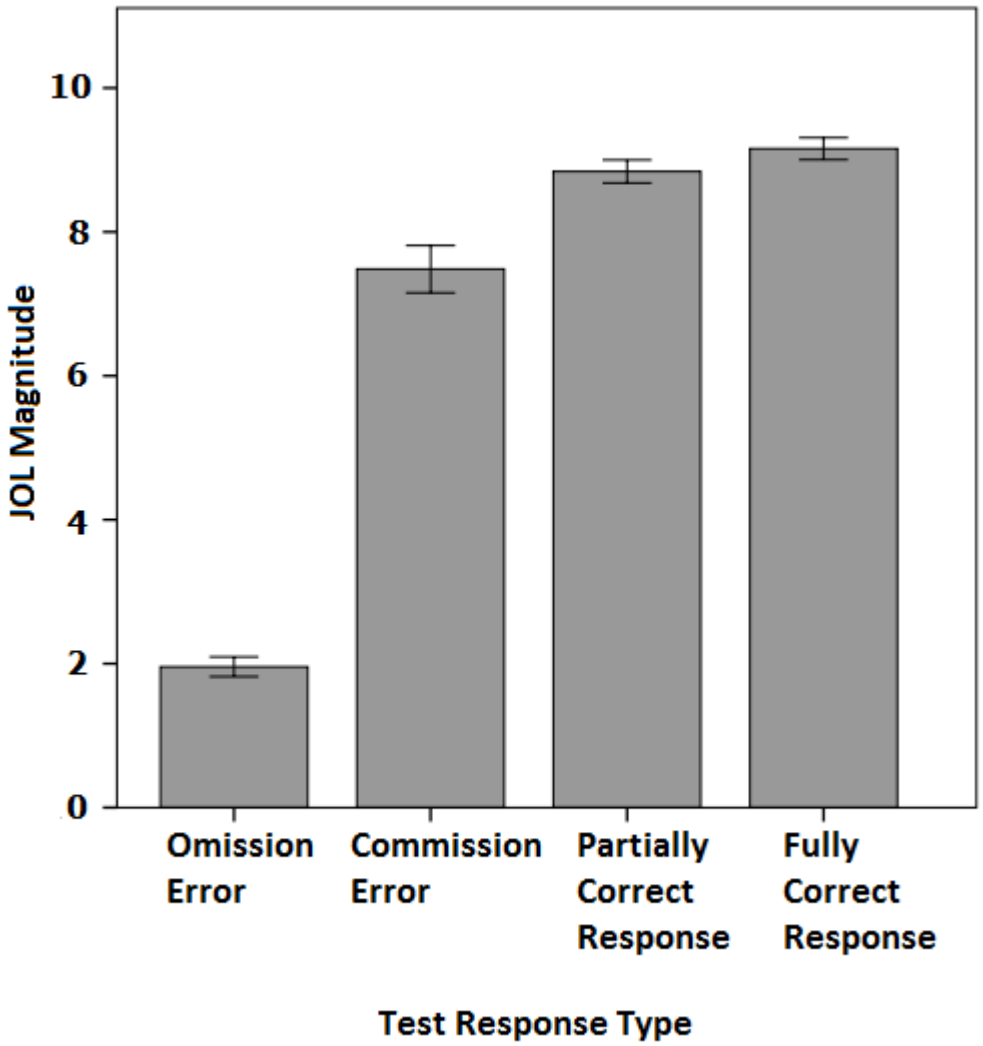
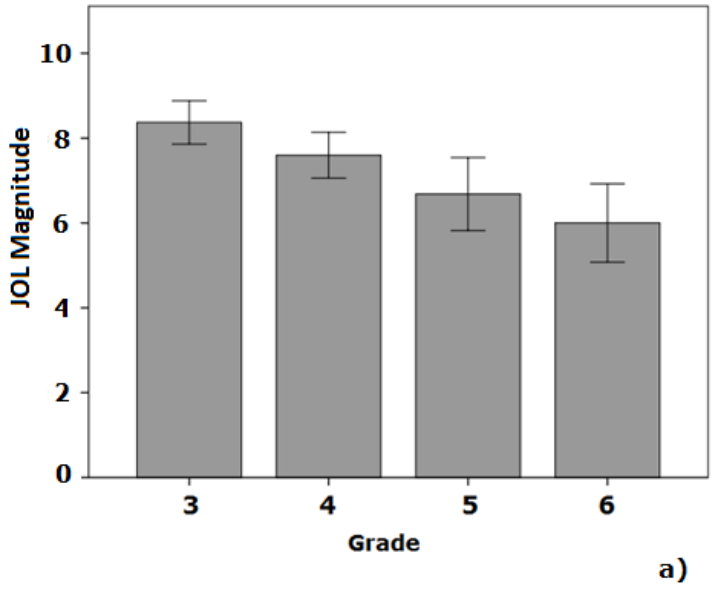


Figure 2

Effects of Grade Level (Grade 3 – 6) on JOLs for commission errors (Figure 2a) and partially correct responses (Figure 2b). Error bars indicate the 95% confidence interval.



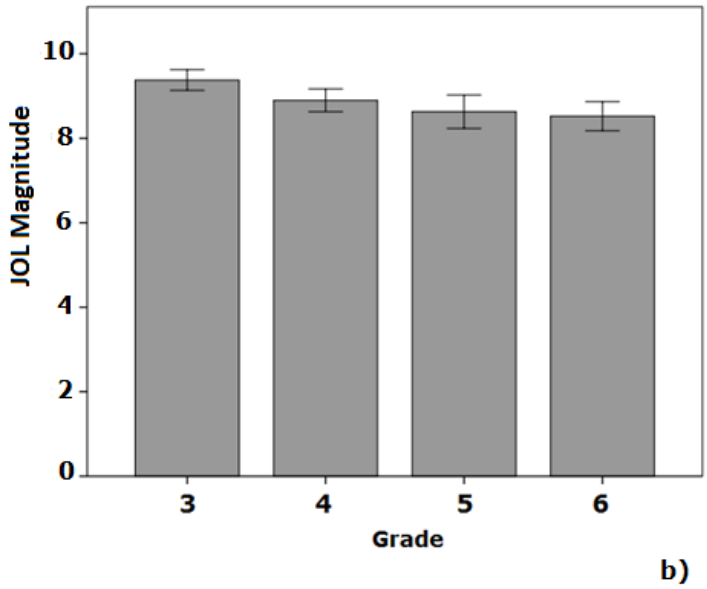


Figure 3

Effects of quality of generated sentences on JOLs. Error bars indicate the 95% confidence interval.

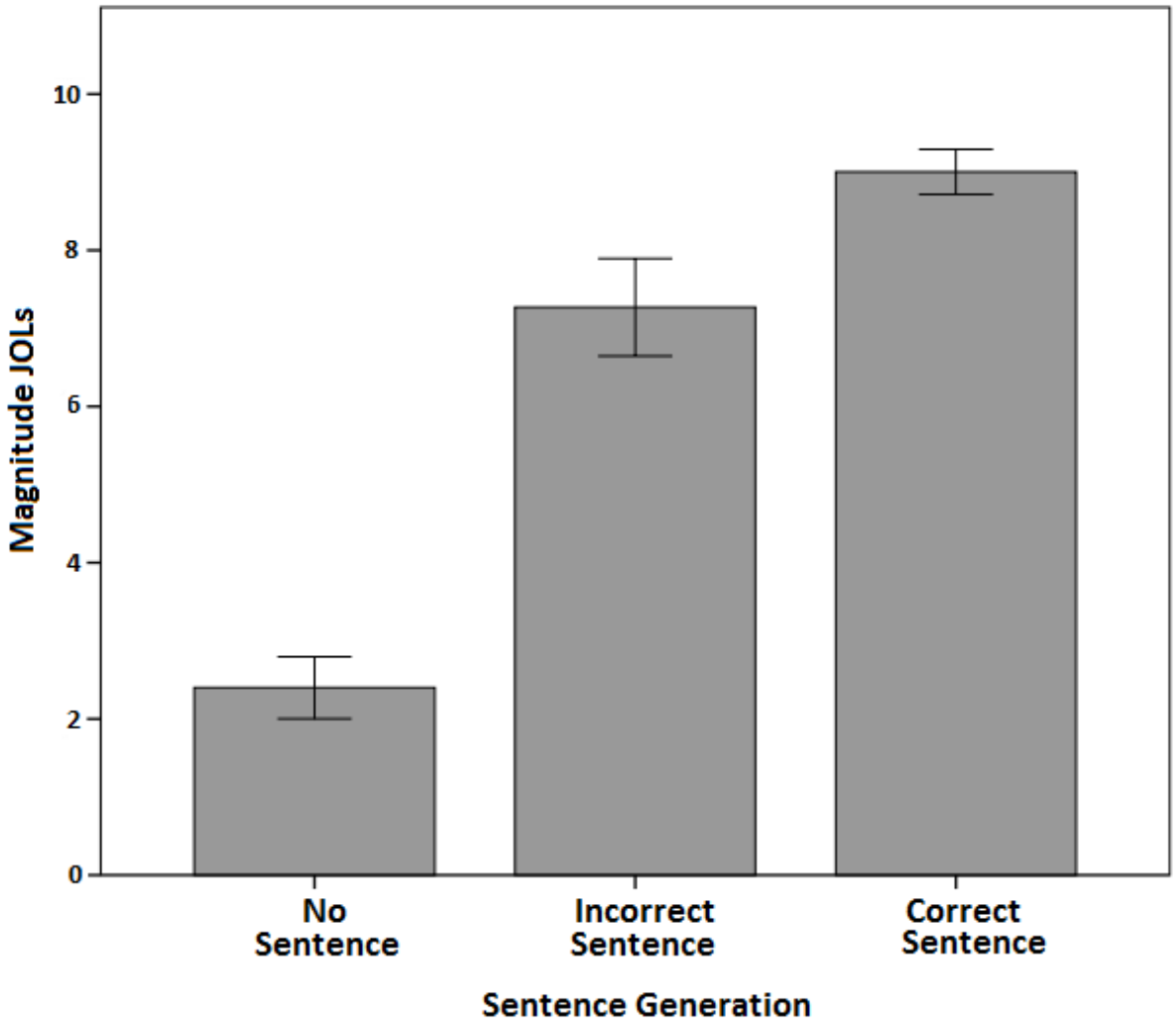


Figure 4

Effects of grade level and sentence generation on JOLs for commission errors and partially correct responses. Error bars indicate the 95% confidence interval.